

A COMBINED EFFICIENT DESIGN FOR BIOMARKER DATA SUBJECT TO A LIMIT OF DETECTION DUE TO MEASURING INSTRUMENT SENSITIVITY¹

BY ENRIQUE F. SCHISTERMAN, ALBERT VEXLER, AIJUN YE
AND NEIL J. PERKINS

*Eunice Kennedy Shriver National Institute of Child Health and Human
Development, University of New York at Buffalo, Eunice Kennedy Shriver
National Institute of Child Health and Human Development and Eunice
Kennedy Shriver National Institute of Child Health and Human
Development*

Pooling specimens, a well-accepted sampling strategy in biomedical research, can be applied to reduce the cost of studying biomarkers. Even if the cost of a single assay is not a major restriction in evaluating biomarkers, pooling can be a powerful design that increases the efficiency of estimation based on data that is censored due to an instrument's lower limit of detection (LLOD). However, there are situations when the pooling design strongly aggravates the detection limit problem. To combine the benefits of pooled assays and individual assays, hybrid designs that involve taking a sample of both pooled and individual specimens have been proposed. We examine the efficiency of these hybrid designs in estimating parameters of two systems subject to a LLOD: (1) normally distributed biomarker with normally distributed measurement error and pooling error; (2) Gamma distributed biomarker with double exponentially distributed measurement error and pooling error. Three-assay design and two-assay design with replicates are applied to estimate the measurement and pooling error. The Maximum likelihood method is used to estimate the parameters. We found that the simple one-pool design, where all assays but one are random individuals and a single pooled assay includes the remaining specimens, under plausible conditions, is very efficient and can be recommended for practical use.

Received December 2010; revised June 2011.

¹Supported by the Intramural Research Program of the Epidemiology branch of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH.

Key words and phrases. Measurement error, pooling, limit of detection, cost-efficient design, three-assay design, two-assay design, duplicate, one-pool design.

This is an electronic reprint of the original article published by the
Institute of Mathematical Statistics in *The Annals of Applied Statistics*,
2011, Vol. 5, No. 4, 2651–2667. This reprint differs from the original in pagination
and typographic detail.

1. Introduction. Epidemiological studies frequently investigate the relationship between biomarkers and disease. In such studies, assaying specimens for biomarkers can be expensive. For example, a single assay to measure polychlorinated biphenyl (PCB) costs between \$500 and \$1,000 [Louis et al. (2005)]. The high cost severely constrains the number of assays that can be performed in a study, thereby limiting the study's ability to characterize a biomarker-disease association.

Two study designs, the pooling design and the simple random sampling design, have been proposed to reduce total assaying cost. Pooling involves assaying only pooled, that is, physically mixed, specimens [Sham et al. (2002)]. Each pooled specimen is obtained by mixing pooling group size p individual specimens together, and each pooled specimen is assumed to contain an amount of biomarker that is the mean of the amounts contained in its constituent individual specimens [Vexler, Liu and Schisterman (2006), Faraggi, Reiser and Schisterman (2003), Schisterman et al. (2001, 2005), Vexler et al. (2008)]. Simple random sampling involves assaying only a simple random sample of individual specimens [Dorfman (1943), Liu and Schisterman (2003), Liu, Schisterman and Teoh (2004), Vexler, Schisterman and Liu (2008), Weinberg and Umbach (1999), Zhang and Gant (2005)].

Not only does cost hinder the characterization of a biomarker-disease association, instrument sensitivity does as well. An instrument may be unable to detect an amount of biomarker below a certain level, the lower limit of detection (LLOD) [Vexler, Liu and Schisterman (2006), Mumford et al. (2006), Vexler et al. (2008), Schisterman et al. (2006)]. Biomarker values above the LLOD are numerically determined, but values below the LLOD are censored. Because instrument sensitivity is an important issue in many areas such as occupational medicine and epidemiology, LLOD issues have been extensively dealt with in the biostatistical literature [Schisterman et al. (2006), Richardson and Ciampi (2003)].

Investigations of the efficiencies of pooling and simple random sampling in parameter estimation when data are subject to a LLOD have been performed. Mumford et al. (2006) and Vexler, Liu and Schisterman (2006) showed that, in the context of biomarker mean and variance estimation, there is always an interval of LLOD values for which pooling is more efficient than simple random sampling and sometimes even more efficient than assaying each and every individual specimen. This phenomenon can be explained by the fact that, when a LLOD is below the mean of a biomarker distribution, a pooled assay has a greater chance of being above the LLOD than an individual assay [Schisterman and Vexler (2008)]. Mumford et al. (2006) also showed that pooling is more efficient than simple random sampling at estimating the area under the receiver operating characteristic curve (AUC) when the LLOD affects less than 50% of the data. However, when the LLOD is substantially greater than the mean of the biomarker distri-

bution, the pooling design is less efficient than simple random sampling at estimating the AUC. Furthermore, the reconstruction of individual assays' characteristics from pooled data is generally a complex issue [Vexler, Schisterman and Liu (2008)].

The merits of pooling and simple random sampling led to the consideration of hybrid designs, which combine pooling and simple random sampling. Some randomly sampled individual specimens are each assayed, and the remaining assays are pooled assays. The efficiency of hybrid designs at parameter estimation has been considered when data are not affected by a LLOD [Schisterman et al. (2010)]. The present article extends previous work by examining the efficiency of a variety of hybrid designs at estimating biomarker distribution parameters and any assaying errors, when assays are affected by a LLOD. When LLOD is present, ignoring missing or replacing missing with a value might lead to severe bias. So it is important to extend our previous work by including LLOD. Furthermore, we demonstrate some hybrid designs under different situations in this article. We consider the efficiency of hybrid designs under various combinations of pooling error and measurement error. Particularly, we are interested in a special case of the general hybrid design, which we call the one-pool design, where all assays but one are random individuals and a single pooled assay includes the remaining specimens. This one-pool design is easy to execute in practice. Our approaches can apply to the upper limit detection (ULOD) as well.

In the following sections we examine the efficiencies of hybrid designs when data are subject to various errors and LLOD. Three-assay design and two-assay design with replicates are applied to account for the pooling error and measurement error. Three-assay design combines one individual sampling group and two pooling groups with different pooling size; while the two-assay design with replicates combines an individual sampling group and one pooling group where each group is measured in replicate. Both designs can be used to estimate the parameters of the biomarker, measurement error and pooling error. The variances of parameters are evaluated for both normally and Gamma distributed biomarker levels. Last, we apply hybrid design to two cases: (1) normally distributed data on cholesterol, a coronary heart disease biomarker and (2) Gamma distributed data on a chemokine biomarker with double exponentially distributed measurement error and pooling error.

2. Pooled-unpooled hybrid design subject to a LLOD. In this section we describe a hybrid design, which combines assays on individual specimens and assays on pooled specimens, when assays are subject to a LLOD. Suppose we have N uncorrelated specimens $\{X_s, s = 1, \dots, N\}$, and we can perform only n assays. Let α be the proportion of n that are assays of individual specimens randomly sampled from all individual specimens. When $\alpha = 1$, only n of the N specimens are used for a simple random sampling

design. We measure αn individual specimens $\{X_s, s = 1, \dots, \alpha n\}$ and use the remaining $N - \alpha n$ individual specimens $\{X_s, s = \alpha n + 1, \dots, N\}$ to create $(1 - \alpha)n$ pooled specimens $\{X_i^{(p)}, i = 1, \dots, (1 - \alpha)n\}$. Here we use subscript i to indicate assays. Ideally we would obtain pooled measurements

$$X_i^{(p)} = \frac{1}{p} \sum_{s=(i-1)p+\alpha n+1}^{ip+\alpha n} X_s,$$

where p is pooling group size, $p = \lceil \frac{N-\alpha n}{(1-\alpha)n} \rceil$. Here $\lceil x \rceil$ is the integer round of a quantity x . When $\alpha n = n - 1$, we have one-pool design with $n - 1$ individual assays $\{X_i, i = 1, \dots, n - 1\}$ and 1 pooled assay $\{X_1^{(p)}\}$. $\alpha_{\text{one-pool}} = 1 - \frac{1}{n}$ is the maximum of α under hybrid design.

In this article we study the hybrid design in a realistic scenario where assays have measurement error and pooling error as well as subject to a LLOD. A simple two-assay hybrid design composed of an individual assay group and a pooled assay group is not enough to estimate both measurement error and pooling error. We can apply two approaches to estimate both errors: (1) three-assay hybrid design and (2) two-assay hybrid design with replicates.

2.1. Three-assay hybrid design. A three-assay hybrid design consists of three different groups, an individual group $Z^{(1)}$, a pooled group $Z^{(p_1)}$ of pooling group size p_1 , and a pooled group $Z^{(p_2)}$ of pooling group size p_2 . Let α be the fraction of assays that are individual assays, and β the fraction of assays that are second pooled assays with pooling size p_2 . The numbers of the assays in each group are $n_1 = \alpha n$, $n_{p_1} = (1 - \alpha - \beta)n$, and $n_{p_2} = \beta n$, respectively. The total number of the specimens are $N = \alpha n + (1 - \alpha - \beta)np_1 + \beta np_2$. Given β and p_2 , we can obtain $p_1 = \lceil \frac{N-\alpha n-\beta np_2}{(1-\alpha-\beta)n} \rceil$. Due to the LLOD, each observation takes the following forms:

$$Z_i^{(w)} = \begin{cases} X_i^{(w)} + \gamma(w)e_i^{(p)} + e_i^{(m)}, & X_i^{(w)} + \gamma(w)e_i^{(p)} + e_i^{(m)} \geq \text{LLOD}, \\ N/A, & X_i^{(w)} + \gamma(w)e_i^{(p)} + e_i^{(m)} < \text{LLOD}, \end{cases}$$

where $w = 1, p_1, p_2$ ($p_1 \neq p_2$, since the three-assay design reduces to the two-assay design when $p_1 = p_2$), $i = 1, \dots, n_w$, $X_i^{(1)}$ are the individual specimens, $e_i^{(m)}$ is measurement error, $e_i^{(p)}$ is pooling error, and $\gamma(w)$ is a known function such that $\gamma(1) = 0$. For simplicity, we assume $\gamma(p_1) = \gamma(p_2) = 1$. When $\beta = 0$, three-assay design reduces to two-assay design. When $\alpha n = n - 1 - \beta n$, we have one-pool design with $n - 1 - \beta n$ individual assays $\{X_i, i = 1, \dots, n - 1 - \beta n\}$, 1 pooled assay $\{X_1^{(p_1)}\}$ with pooling size p_1 , and βn pooled assays $\{X_i^{(p_2)}, i = 1, \dots, \beta n\}$ with pooling size p_2 . We have $\alpha_{\text{one-pool}} = 1 - \beta - \frac{1}{n}$. When $\beta = 0$, three-assay design reduces to two-assay design, that is, $\alpha_{\text{one-pool}} = 1 - \frac{1}{n}$.

2.2. *Two-assay design with replicates.* Another approach to estimate pooling and measurement errors is two-assay design with replicates. In practice, laboratories often measure the assays twice. When a specimen is measured twice, for individual samples, we have

$$Z_{i1}^{(1)} = X_i + e_{i1}^{(m)}, \quad Z_{i2}^{(1)} = X_i + e_{i2}^{(m)},$$

where $Z_{i1}^{(1)}$ and $Z_{i2}^{(1)}$ are measured values, X is the true value, and $e_{i1}^{(m)}$ and $e_{i2}^{(m)}$ are measurement errors. In practice, laboratories often use the average of $Z_{i1}^{(1)}$ and $Z_{i2}^{(1)}$ as the true biomarker value. We also have

$$(1) \quad \Delta Z_i^{(1)} = Z_{i1}^{(1)} - Z_{i2}^{(1)} = e_{i1}^{(m)} - e_{i2}^{(m)}.$$

By fitting the distribution of $\Delta Z_i^{(1)}$, we can obtain the parameter for measurement error $e^{(m)}$. For pooled assays, we have

$$Z_{i1}^{(p)} = X_i + e_1^{(m)} + e_1^{(p)}, \quad Z_{i2}^{(p)} = X_i + e_{i2}^{(m)} + e_{i2}^{(p)},$$

where $e_1^{(p)}$ and $e_{i2}^{(p)}$ are pooling errors. We also have

$$(2) \quad \Delta Z^{(p)} = Z_{i1}^{(p)} - Z_{i2}^{(p)} = (e_1^{(m)} + e_1^{(p)}) - (e_{i2}^{(m)} + e_{i2}^{(p)}).$$

By fitting the distribution of $\Delta Z_i^{(p)}$, we can obtain the parameter for the sum of measurement error and pooling error $e^{(m)} + e^{(p)}$. After we obtain the estimates of the pooling and measurement errors, we can use a two-assay design involving one individual sampling group and only one pooling group to estimate the parameters of the biomarker.

2.3. *Maximum likelihood estimate.* The literature on limit of detection is largely maximum likelihood (ML) due to a need to assume a distribution for the data that are unmeasurable below the limit of detection. For insight below the limit of detection, the distribution above is assessed and assumed consistent below. ML estimation follows naturally after this. One simple way to address the LLOD is to substitute a replacement value for unobservable data. However, it will lead to biased assessment and it has been shown that the best value is often $E[X|X < d]$ and required the same assumption on the distribution below the limit of detection. In this article, we use the ML method to handle LLOD data because it yields asymptotically unbiased estimates of the parameters [Gupta (1952), Chapman (1956)]. We consider two cases: (1) normally distributed biomarker with normally distributed measurement error and pooling error, and (2) Gamma distributed biomarker with double exponentially distributed measurement error and pooling error.

2.3.1. *Normal distributed biomarker and errors.* Let the individual biomarker values be independently and identically distributed as follows:

$$X_i \sim N(\mu_x, \sigma_x^2), \quad i = 1, \dots, \alpha n,$$

where $\alpha \in (0, 1)$. By applying the pooling design based on $N - \alpha n$ assays, ideally we would obtain pooled measurements following normal distribution

$$X_i^{(p)} \sim N\left(\mu_x, \frac{\sigma_x^2}{p}\right), \quad i = 1, \dots, (1 - \alpha)n.$$

We assume that the measurement error $e^{(m)}$ and pooling error $e^{(p)}$ also follow independent normal distribution

$$e_i^{(m)} \sim N(0, \sigma_m^2), \quad e_i^{(p)} \sim N(0, \sigma_p^2), \quad i = 1, \dots, n.$$

The detailed likelihood function is available in the supplementary material [Schisterman et al. (2011), Section 1].

2.3.2. Gamma distributed biomarker and double exponentially distributed errors. In certain situations, the distribution of the biomarker values is skewed, and the normality assumptions cannot be applied. In these circumstances, the Gamma distribution is a reasonable alternative. Furthermore, the distribution of measurement and pooling errors can vary by shape, and the normality assumptions are not always reasonable. In these cases, double exponential distribution might be appropriate, because it is symmetric and mean zero. Suppose that the individual biomarker X_i follows a Gamma distribution

$$X_i \sim g(x; a, b) = \frac{1}{b^a \Gamma(a)} e^{-x/b} x^{a-1}, \quad i = 1, \dots, \alpha n.$$

For pooled assays with pooling size p , using the additive property of the Gamma distribution, we have

$$X_i^{(p)} \sim g(x; ap, b/p), \quad i = 1, \dots, (1 - \alpha)n,$$

and the measurement error and pooling error follow a double exponential distribution with scale parameters c and d , respectively,

$$e_i^{(m)} \sim h(x; c) = \frac{1}{2c} e^{-|x|/c}, \quad e_i^{(p)} \sim h(x; d) = \frac{1}{2d} e^{-|x|/d}, \quad i = 1, \dots, n.$$

The detailed likelihood function is available in the supplementary material [Schisterman et al. (2011), Section 2].

2.4. Evaluation. In this section we evaluate three cases: (1) normally distributed biomarker with negligible measurement error and pooling error under two-assay design, (2) normally distributed biomarker with normally distributed measurement error and pooling error under three-assay design, and (3) Gamma distributed biomarker and double exponentially distributed measurement error and pooling error under two-assay design.

2.4.1. Normal case with negligible pooling error and measurement error.

We are interested in the one-pool design, a special case of the hybrid design, because it is simple and easily executed in practice. The one-pool design fixes the $\alpha n = n - 1$ individual sampling group, leaving $(1 - \alpha)n = 1$ of the remaining $N - (n - 1)$ specimens. We first use a simple case with negligible pooling error and measurement error to illustrate the efficiency of the one-pool design.

When random sampling and pooling are combined in the hybrid design, the data consist of individual and pooled observations $\{Z_1^{(1)}, \dots, Z_{[\alpha n]}^{(1)}, Z_1^{(p)}, \dots, Z_{[(1-\alpha)n]}^{(p)}\}$. If we assume that the measurement error and pooling error are negligible, that is, $e^{(m)} = 0$ and $e^{(p)} = 0$, the three-assay design is reduced to a two-assay design ($\beta = 0$). Each observation takes the form

$$Z_i^{(w)} = \begin{cases} X_i^{(w)}, & X_i^{(w)} \geq \text{LLOD}, \\ N/A, & X_i^{(w)} < \text{LLOD}. \end{cases}$$

The log-likelihood function for normal distribution is a function of only parameters μ_x and σ_x . To calculate the MLEs of parameters μ_x and σ_x , we solve the system of log-likelihood first derivative equations $\{\frac{\partial \ell}{\partial \mu_x} = 0, \frac{\partial \ell}{\partial \sigma_x} = 0\}$. Expressions for the log-likelihood equations and the entries of Fisher information matrix I can be found in the supplementary material [Schisterman et al. (2011), Section 3]. The asymptotic variances of the estimators can be analyzed with respect to α (the proportion of assays that are individual assays), and an α that minimizes the variance of an MLE can be proposed.

Figure 1 illustrates the asymptotic variances $\text{Var}(\hat{\mu}_x)$ and $\text{Var}(\hat{\sigma}_x)$ versus α for LLOD = -5, -0.5, -0.1, 0, 0.01, 0.04, 0.1, 0.3 and 0.5 from bottom to top with $N = 1,000$, $n = 100$, $\mu_x = 0$ and $\sigma_x = 1$. Note that the rightmost point is at $\alpha = (n - 1)/n$, that is, one-pool design, rather than $\alpha = 1$.

When LLOD is negligible (e.g., LLOD = -5), $\text{Var}(\hat{\mu}_x)$ is approximately constant for $\alpha < 1$ in Figure 1(a). $\text{Var}(\hat{\mu}_x)$ increases with the increase of LLOD. For $\text{LLOD} \leq \mu_x$, $\text{Var}(\mu_x)$ decreases as α increases, for example, $\text{LLOD} = \mu_x$ (i.e., 0) and $\mu_x - 0.1\sigma_x$ (i.e., -0.1). $\text{Var}(\mu_x)$ takes the minimum at $\alpha_{\text{one-pool}} = (n - 1)/n = 0.99$. When $\text{LLOD} > \mu_x$, $\text{Var}(\hat{\mu}_x)$ takes a minimum value at an $0 < \alpha < \alpha_{\text{one-pool}}$ as shown in Figure 1(a) and (b). A hybrid design is more efficient than only measuring pooled assays or only measuring individual assays. When $\text{LLOD} < \mu_x$, the traditional pooling design ($\alpha = 0$) is more efficient than simple random sampling [Vexler, Liu and Schisterman (2006), Mumford et al. (2006)]. However, when a pooled-unpooled hybrid design is applicable, when $\text{LLOD} \leq \mu_x$ and the objective is the estimate μ_x , we recommend a one-pool design given that pooling and measurement errors are negligible. However, when N is very large, pooling $N - (n - 1)$ specimens might exceed the laboratory limitations.

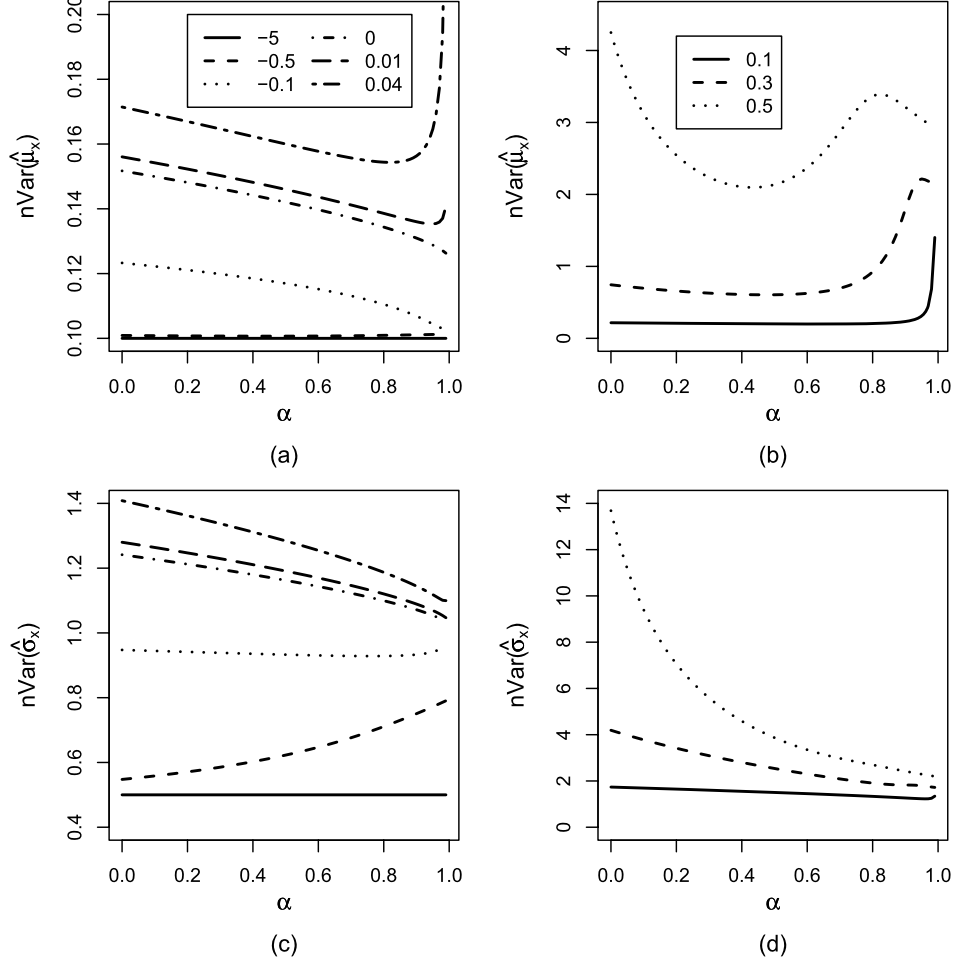


FIG. 1. $n\text{Var}(\hat{\mu}_x)$ and $n\text{Var}(\hat{\sigma}_x)$ versus the proportion of individual assays to the measured assays α in the absence of measurement and pooling errors with $\text{LLOD} = -5, -0.5, -0.1, 0, 0.01, 0.04, 0.1, 0.3$ and 0.5 from bottom to top; $N = 1,000$, $n = 100$, $\mu_x = 0$ and $\sigma_x = 1$.

Figure 1(c) shows $\text{Var}(\hat{\sigma}_x)$ is approximately constant as well when LLOD is absent (e.g., $\text{LLOD} = -5$). For $\text{LLOD} < \mu_x$ (e.g., $\text{LLOD} = \mu_x - 0.5\sigma_x$), pool design ($\alpha = 0$) minimizes $\text{Var}(\hat{\sigma}_x)$. For $\text{LLOD} \geq \mu_x$ (e.g., $0, 0.01, 0.04, 0.1, 0.3$ and 0.5), $\text{Var}(\hat{\sigma}_x)$ takes the minimum when the one-pool design is used, as shown in Figure 1(c) and (d).

The traditional pooling design involves obtaining n pooled assays with pooling group size $p = N/n$. With this design, the variance of the μ_x -estimator based on n measurements of the pooled assays is σ_x^2/N . For one-pool design with pooling group size $p = N - n + 1$, when the LLOD is not in

effect, the MLE of μ_x based on the combined data $\{Z_1^{(1)}, \dots, Z_{n-1}^{(1)}, Z_1^{(N-n+1)}\}$ is the following:

$$\begin{aligned}\hat{\mu}_x &= \frac{1}{n-1+p} \left(\sum_{i=1}^{n-1} Z_i^{(1)} + p Z_1^{(p)} \right) \\ &= \frac{1}{N} \left\{ \sum_{s=1}^{n-1} X_s + (N-n+1) \left(\sum_{s=n}^N X_s / (N-n+1) \right) \right\}.\end{aligned}$$

Thus, the one-pool design $\{Z_1^{(1)}, \dots, Z_{n-1}^{(1)}, Z_1^{(N-n+1)}\}$ allows estimation of μ_x . $\text{Var}(\hat{\mu}_x)$ is equivalent to that based on traditionally pooled data $\{Z_1^{(N/n)}, \dots, Z_n^{(N/n)}\}$. This variance is not equivalent to that based on a simple random sample of individual assays $\{Z_1^{(1)}, \dots, Z_n^{(1)}\}$. The same conclusion can be shown regarding the σ_x^2 -estimation. This proposed one-pool design is easier to execute than traditional pooling. Moreover, if the parametric assumptions regarding the sample distribution are rejected, the data $\{Z_1^{(1)}, \dots, Z_{n-1}^{(1)}, Z_1^{(N-n+1)}\}$ can easily be used to estimate the unknown distribution, whereas reconstruction of the distribution function of X based on $\{Z_1^{(N/n)}, \dots, Z_n^{(N/n)}\}$ is a very complicated problem [Vexler, Schisterman and Liu (2008)]. Even when the LLOD has a role, namely, when $\text{LLOD} \leq \mu_x$, as in Figure 1(a), we can suggest the simple one-pool design.

2.4.2. Normal case with nonnegligible measurement error and pooling error. When pooling error and measurement error are nonnegligible, one approach to estimating the pooling and measurement errors is a three-assay design, as mentioned at the beginning of this section. The expressions for the normally distributed log-likelihood equations and the entries of Fisher information matrix I can be found in the supplementary material [Schisterman et al. (2011), Section 4]. Figure 2 depicts the evolutions of $n \text{Var}(\hat{\mu}_x)$, $n \text{Var}(\hat{\sigma}_x)$, $n \text{Var}(\hat{\sigma}_p)$ and $n \text{Var}(\hat{\sigma}_m)$ with $N = 1,000$, $n = 100$, $\sigma_x = 1$, $\sigma_p = 0.3$ and $\sigma_m = 0.4$. The curves from bottom to top are for $\text{LLOD} = -5, -0.5, 0$ and 0.5 , respectively. Because our hybrid design involves two pooling groups, we set the proportion of the second pooling group $\beta = 0.4$ and pooling size $p_2 = 5$. Note that the rightmost point $\alpha = [(1-\beta)n - 1]/n = 0.59$ is corresponding to the one-pool design that consists of $(1-\beta)n - 1 = 59$ individual assays, 1 pooled assay with pooling size $p_1 = 741$, and $\beta n = 40$ pooled assays with pooling size $p_2 = 5$.

As LLOD increases, $\text{Var}(\hat{\mu}_x)$, $\text{Var}(\hat{\sigma}_x)$, $\text{Var}(\hat{\sigma}_m)$ and $\text{Var}(\hat{\sigma}_p)$ increase. $\text{Var}(\hat{\mu}_x)$ increases as α increases, that is, the pooled design minimizes $\text{Var}(\hat{\mu}_x)$. $\text{Var}(\hat{\sigma}_x)$, $\text{Var}(\hat{\sigma}_m)$ and $\text{Var}(\hat{\sigma}_p)$ obtain the minimum under the hybrid design. We provide R code as the supplementary material [Schisterman et al. (2011)] to calculate $\text{Var}(\hat{\mu}_x)$, $\text{Var}(\hat{\sigma}_x)$, $\text{Var}(\hat{\sigma}_m)$ and $\text{Var}(\hat{\sigma}_p)$.

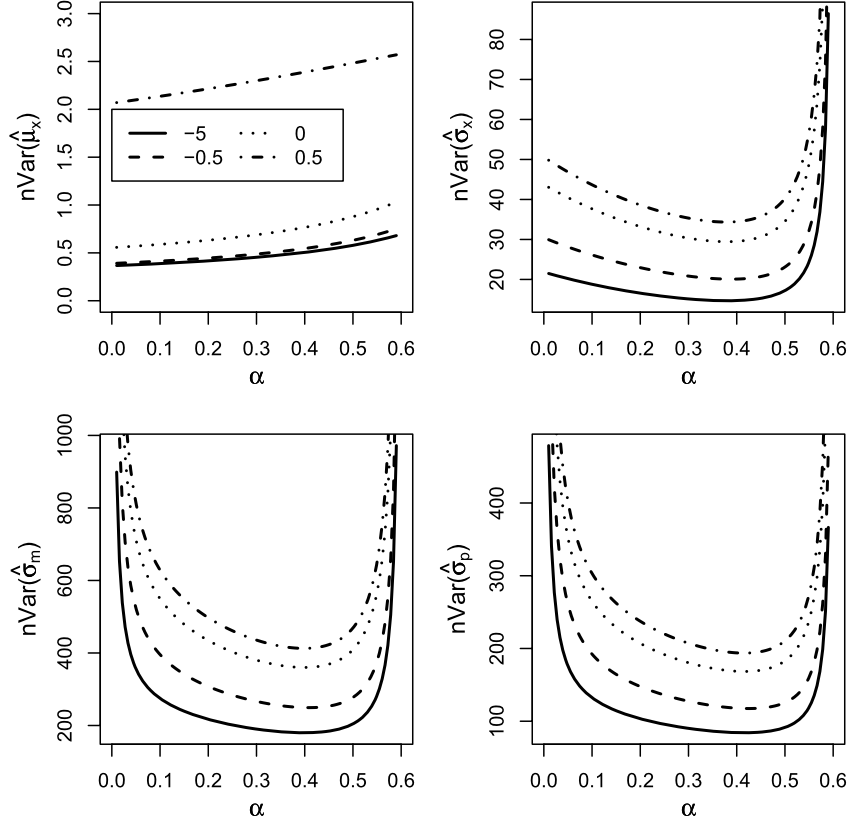


FIG. 2. $n \text{Var}(\hat{\mu}_x)$, $n \text{Var}(\hat{\sigma}_x)$, $n \text{Var}(\hat{\sigma}_m)$, and $n \text{Var}(\hat{\sigma}_p)$ versus the proportion of individual assays to the measured assays α in the presence of measurement and pooling errors under three-assay design with LLOD = $-5, -0.5, 0$ and 0.5 from bottom to top; the proportion of the second pooled assays $\beta = 0.4$, pooling size $p_2 = 5$, $N = 1,000$, $n = 100$, $\mu_x = 0$, $\sigma_x = 1$, $\sigma_m = 0.3$ and $\sigma_p = 0.4$.

2.4.3. Gamma case with measurement error and pooling error. In this subsection we study the situation with Gamma distributed biomarker, double exponentially distributed measurement error and pooling error by Monte Carlo simulation. Two-assay design can be used when we know the variances of measurement error and pooling error. The parameters for the Gamma distributed biomarker are $a = 1.5$ and $b = 0.1$. So the mean of the individual biomarker is $E(X) = ab = 0.15$, and the variance $\text{Var}(X) = ab^2 = 0.015$. The parameters for double exponentially distributed measurement error and pooling error are $c = 0.02$ and $d = 0.03$, respectively. Both errors are mean zero and the variance of measurement error is $\text{Var}(e^{(m)}) = 2c^2 = 0.0008$, and the variance of pooling error $\text{Var}(e^{(p)}) = 2d^2 = 0.00018$. The number of specimens is $N = 1,000$ and the number of assays is $n = 100$. 1,000 simulations

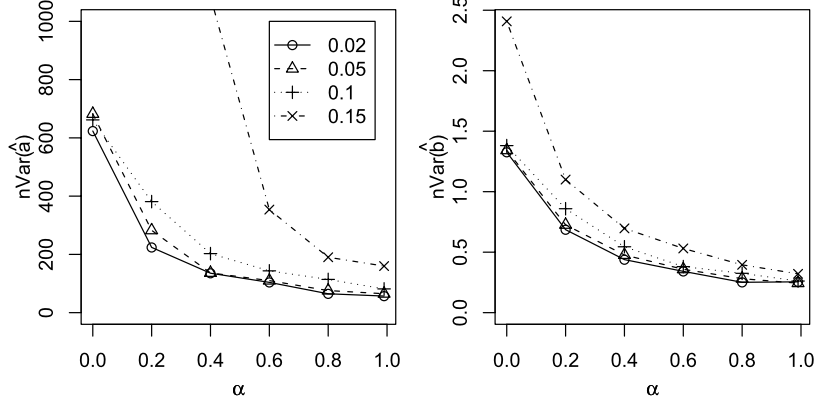


FIG. 3. $n\text{Var}(\hat{a})$ and $n\text{Var}(\hat{b})$ versus the proportion of individual assays to the measured assays α for simulated Gamma distributed biomarkers with double exponentially distributed measurement and pooling errors with $N = 1,000$, $n = 100$, $a = 1.5$, $b = 0.1$, $c = 0.02$, $d = 0.03$, $LLOD = 0.02, 0.05, 0.1$ and 0.15 .

were performed to evaluate $\text{Var}(\hat{a})$ and $\text{Var}(\hat{b})$ at $\alpha = 0, 0.2, 0.4, 0.6, 0.8$ and 0.99 , subject to $LLOD = 0.02, 0.05, 0.1$ and 0.15 .

The simulation results are presented in Figure 3. $\text{Var}(\hat{a})$ and $\text{Var}(\hat{b})$ increases with the increase of $LLOD$. Both $\text{Var}(\hat{a})$ and $\text{Var}(\hat{b})$ decrease with the increase of α . They are minimized under the one-pool design ($\alpha = 0.99$). When $LLOD < E(X)$, $\text{Var}(\hat{a})$ does not change much with the increase of α . However, when $LLOD = E(X)$, $\text{Var}(\hat{a})$ becomes significantly larger, especially when α is small. It is five-fold larger than with other $LLOD$ values for pool design ($\alpha = 0$). $\text{Bias}(\hat{a})$ and $\text{Bias}(\hat{b})$ for finite sample size are presented in Section 5 of the supplementary material [Schisterman et al. (2011)]. They are relatively small except for large $LLOD$, for example, $LLOD = 0.15$ (61% missing for individual sampling).

3. Application.

3.1. Normally distributed biomarker with negligible measurement and pooling errors. In order to investigate the efficiency of the hybrid design, we bootstrapped by using real data from a study of biomarkers of coronary heart disease. In this study, cholesterol level, a biomarker for coronary heart disease, was measured for 40 individuals that had a normal rest electrocardiogram, were free of symptoms, and had no previous cardiovascular procedures or myocardial infarctions. The mean of the individual biomarker assays is 205.53 mg/dl and the standard deviation is 42.29 mg/dl. The Shapiro-Wilk test for normality suggests that the individual assays follow a normal distribution.

We assume that we have $N = 40$ specimens, we can only afford to perform $n = 20$ assays, and the measurement error and pooling error are negligible.

TABLE 1
Parameters used for normally distributed biomarker ignoring errors with number of samples $N = 40$ and the number of assays $n = 20$

α	0	0.5	0.75	0.8	0.9	0.95
Number of individual assays	0	10	15	16	18	19
Number of pooled assays	20	10	5	4	2	1
Pooling size p	2	3	5	6	11	21

Artificial LLOD = 0, 150, 170, 180, 200, 205 and 210 are applied to the cholesterol data. We evaluated six designs, involving α values from Table 1. The rightmost one ($\alpha = 0.95$) is a one-pool design. To generate the pooled data with different pooling size p , we pooled the individual assays together, and used the average values as the measured values of the pooled assays. Then we combined the unpooled and simulated pooled data, and applied the methodology for two-assay design with negligible measurement and pooling error case in Section 2.4.1 to calculate the maximum likelihood estimate of μ_x . This procedure is repeated 100,000 times to obtain $\text{Var}(\hat{\mu}_x)$.

The results are shown in Figure 4. When $\text{LLOD} < \hat{\mu}_x - \hat{\sigma}_x$ (e.g., 0 and 150), $\text{Var}(\hat{\mu}_x)$ is approximately a constant. When $\hat{\mu}_x - \hat{\sigma}_x < \text{LLOD} < \hat{\mu}_x$ (e.g., 170 and 180), $\text{Var}(\hat{\mu}_x)$ decreases as α increases. The minimum is obtained under the one-pool design. When LLOD is close to $\hat{\mu}_x$ (e.g., 200 and 205), $\text{Var}(\hat{\mu}_x)$ takes the minimum at $0 < \alpha < 1$. A hybrid design is favorable. Although the one-pool design does not give the minimum, $\text{Var}(\hat{\mu}_x)$ for the one-pool design (78.2 for LLOD = 200) is close to the minimum (68.7). Due to the simplicity of design, one-pool design can be recommended. When $\text{LLOD} > \hat{\mu}_x$ (e.g., 210), $\text{Var}(\hat{\mu}_x)$ increases as α increases. The maximum of $\text{Var}(\hat{\mu}_x)$ is obtained under one-pool design.

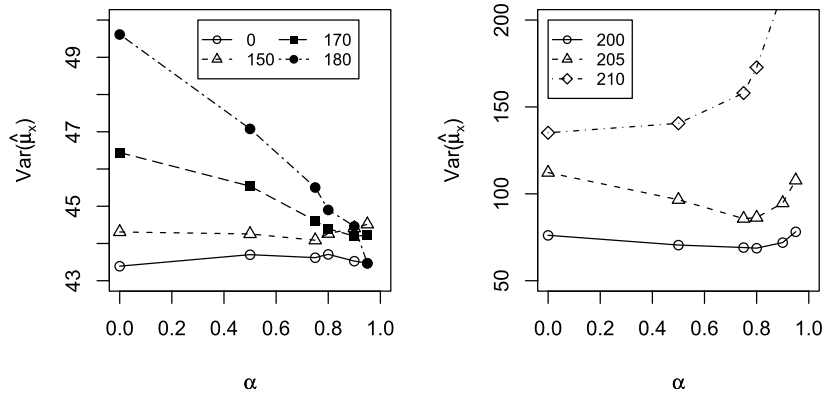


FIG. 4. $\text{Var}(\hat{\mu}_x)$ versus the proportion of individual assays to the measured assays α by bootstrapping with $N = 40$, $n = 20$, LLOD = 0, 150, 170, 180, 200, 205 and 210.

3.2. Gamma distributed biomarker with double exponentially distributed measurement error and pooling error. In this subsection we exemplified the two-assay design with replicates using real data from a study of chemokine biomarker monocyte chemotactic protein-1 (MCP-1). MCP-1 plays a role in a variety of pathological conditions such as inflammatory and immune reactions. Assays are measured in different plates. Each plate has its own LLOD. In this article we use only the data from the plates with LLOD = 0.016, because our model requires the same LLOD. Each plate was measured twice. There are 99 individual sampling assays, and 45 pooled assays with $p = 2$. The mean of the individual sampling assays is 0.189, and the standard deviation is 0.183. The measurement errors can be calculated by the difference of individual sampling assays [see (1)], and the pooling errors can be calculated by the difference of pooled assays; see (2). We used the R package VGAM [Yee (2010)] to fit the difference of individual replicates $\Delta Z^{(1)}$ to obtain the estimate of parameter c . Then we fit the difference of pooled replicates $\Delta Z^{(p)}$, which follows a double exponential distribution with parameter c . The estimated variances of measurement error and pooling error can be obtained by

$$\widehat{\text{Var}}(e^{(m)}) = \frac{\widehat{\text{Var}}(\Delta Z^{(1)})}{2},$$

$$\widehat{\text{Var}}(e^{(p)}) = \frac{\widehat{\text{Var}}(\Delta Z^{(p)}) - \widehat{\text{Var}}(\Delta Z^{(1)})}{2}.$$

After we obtained the estimates of the variances of pooling error and measurement error, we used one individual sampling group and one pooling group to estimate the other parameters, for example, a and b of the Gamma distributed biomarker.

The histograms of individual biomarker $Z^{(1)}$, difference of measurement error $e_1^{(m)} - e_2^{(m)}$ and difference of the sum of measurement error and pooling error $(e_1^{(m)} + e_1^{(p)}) - (e_2^{(m)} + e_2^{(p)})$ are illustrated in Figure 5. The fitting curves are generated by the parameters estimated by the R package VGAM. The estimated parameters are presented in Table 2. For double exponential distribution, the estimated variance is $2s^2$, where s is the scale parameter of double exponential distribution. The estimated $\widehat{\text{Var}}(e^{(m)})$ and $\widehat{\text{Var}}(e^{(p)})$ are presented in Table 2 as well as their corresponding scale parameters. For Gamma distribution, the estimated mean is ab and the estimated variance is ab^2 . Table 2 shows that the sample variances are very close to the estimated variances. The fitting curves in Figure 5 fit the histogram quite well.

For fixed N and n , we need to vary the pooling size p to vary α . However, we only have individual unpooled data and pooled data with pooling size $p = 2$. So we pool the $p = 2$ pooled assays together to generate the data with different pooling size. Because we want to include the measurement error and pooling error in the pooled assays, we used pooled assays rather than

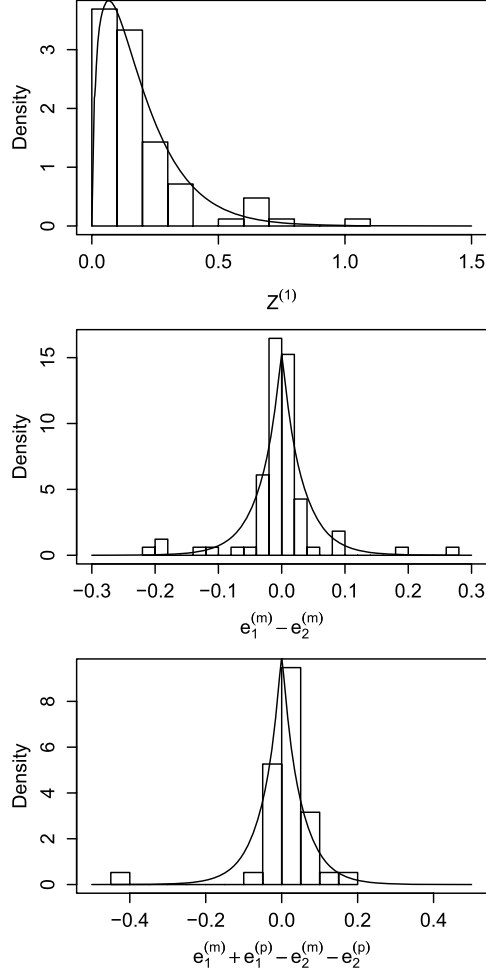


FIG. 5. Histograms of individual biomarker $Z^{(1)}$, difference of measurement error $e_1^{(m)} - e_{i2}^{(m)}$ and difference of the sum of measurement error and pooling error $(e_1^{(m)} + e_1^{(p)}) - (e_{i2}^{(m)} + e_{i2}^{(p)})$.

individual sampling assays to generate pooled assays with different pooling size. For example,

$$\begin{aligned}
 Z^{(p=4)} &= \frac{1}{2}(Z_{i1}^{(p=2)} + Z_{i2}^{(p=2)}) \\
 &= \frac{1}{2}\left(\frac{X_1 + X_2}{2} + e_1^{(p)} + e_1^{(m)} + \frac{X_3 + X_4}{2} + e_{i2}^{(p)} + e_{i2}^{(m)}\right) \\
 &= \frac{1}{4}(X_1 + X_2 + X_3 + X_4) + \frac{1}{2}(e_1^{(p)} + e_{i2}^{(p)}) + \frac{1}{2}(e_1^{(m)} + e_{i2}^{(m)}).
 \end{aligned}$$

TABLE 2

The estimates of the parameters for individual biomarker $Z^{(1)}$, difference of measurement error $e_1^{(m)} - e_{i2}^{(m)}$, difference of the sum of measurement error and pooling error $(e_1^{(m)} + e_1^{(p)}) - (e_{i2}^{(m)} + e_{i2}^{(p)})$, measurement error $e^{(m)}$ and pooling error $e^{(p)}$. Here a and b are the shape and scale parameters of the Gamma distribution, respectively, s are the scale parameters of double exponential distribution

	a	b	s	Mean		Variance	
				Estimated	Sample	Estimated	Sample
$Z^{(1)}$	1.54	0.12		0.189	0.189	0.023	0.034
$e_1^{(m)} - e_{i2}^{(m)}$			0.033	0	-0.0034	0.0022	0.0029
$e_1^{(m)} + e_1^{(p)} - e_{i2}^{(m)} - e_{i2}^{(p)}$			0.050	0	0.012	0.0051	0.0059
$e^{(m)}$			0.023			0.0011	
$e^{(p)}$			0.027			0.0015	

Then we combined individual unpooled data, and measured ($p = 2$) or simulated ($p > 2$) pooled data to generate a hybrid design. The pooling sizes we used are presented in Table 3. We assume that we have $N = 79$ or 80 specimens, and can only afford to perform $n = 40$ assays. Besides the true $\text{LLOD} = 0.016$, additional $\text{LLOD} = 0.05, 0.1$ and 0.15 are applied to evaluate the influence of LLOD .

The results are illustrated in Figure 6. As α increases, $\text{Var}(\hat{a})$ increases then decreases at the one-pool design ($\alpha = 0.975$). One-pool design gives the second minimum. This tendency is different from the simulation result, where $\text{Var}(\hat{a})$ decreases as α increases, and the minimum is reached under one-pool design. When LLOD is very small (i.e., $\text{LLOD} = 0.016$), $\text{Var}(\hat{a})$ does not change much. $\text{Var}(\hat{b})$ decreases as α increases, which is consistent with the simulation result.

4. Summary and discussion. Although the pooling design can increase the efficiency of estimation from data subject to a LLOD , there are situations when the pooling design strongly aggravates the detection limit problem.

TABLE 3

Parameters used for the Gamma distributed biomarker with double exponentially distributed errors and the number of assays $n = 40$

α	0	0.675	0.8	0.925	0.975
Number of individual assays	0	27	32	37	39
Number of pooled assays	40	13	8	3	1
Pooling size p	2	4	6	14	40
Number of samples N	80	79	80	79	79

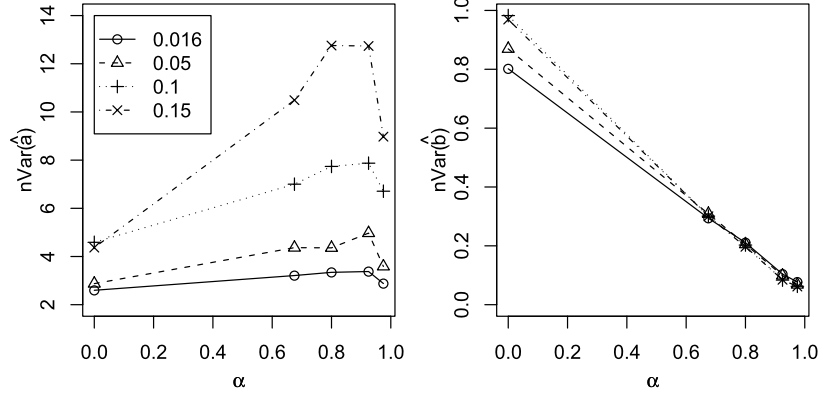


FIG. 6. $n \text{Var}(\hat{a})$ and $n \text{Var}(\hat{b})$ versus the proportion of individual assays to the measured assays α by bootstrapping with $N = 79$ or 80 , $n = 40$, $\text{LOD} = 0.016, 0.05, 0.1$ and 0.15 .

A hybrid design was proposed in order to gain benefits from both individual assays and pooled assays [Schisterman et al. (2010)].

In this article we present methodology for determining a hybrid design that most efficiently estimates parameters from data subject to measurement error, pooling error and a limit of detection. Efficiency is gauged by the variance of a maximum likelihood estimator of a parameter. We demonstrated the asymptotic MLE variances as functions of the proportion of individual assays to the measured assays. To estimate both measurement error and pooling error, a three-assay design or a two-assay design with replicates is needed. We examined two cases: one is with the normally distributed biomarker and errors, the other is with the Gamma distributed biomarker and double exponentially distributed errors.

Under the condition that we have N specimens and we can only perform $n < N$ assays, we evaluated the efficiency of the one-pool hybrid design, which involves assaying $n - 1$ individual specimens and one pooled sample of the remaining $N - (n - 1)$ individual specimens. When measurement error and pooling error are negligible, for the normally distributed biomarker, one-pool design minimizes $\text{Var}(\hat{\mu}_x)$ for $\text{LLOD} \leq \mu_x$ and $\text{Var}(\hat{\sigma}_x)$ for $\text{LLOD} > \mu_x$. When measurement error and pooling error are in effect, the pooled design minimizes $\text{Var}(\hat{\mu}_x)$, while the hybrid design minimize $\text{Var}(\hat{\sigma}_x)$, $\text{Var}(\hat{\sigma}_m)$ and $\text{Var}(\hat{\sigma}_p)$. The α value corresponding to the minimum can be obtained by the R code that we provided as the supplementary material [Schisterman et al. (2011)]. Note that, in practice, our interest is in μ_x , σ_x , and not in σ_p or σ_m . The simulation result shows that it minimizes both $\text{Var}(\hat{a})$ and $\text{Var}(\hat{b})$ for Gamma distribution under complex measurement error and pooling error assumptions. Hence, under the circumstances described above, when one seeks to avoid more complicated procedures for determining and

executing a potentially more efficient hybrid design, the one-pool hybrid design is an efficient and easily implemented alternative to a simple random sample of individual assays.

Acknowledgments. The authors thank the Editor, reviewers, Qian Zhang and Sonya Dasharathy for their valuable comments, and Dr. Brian Whitcomb for the chemokines data.

SUPPLEMENTARY MATERIAL

R code and detailed derivations (DOI: [10.1214/11-AOAS490SUPP](https://doi.org/10.1214/11-AOAS490SUPP); .pdf). R code used to calculate $n \text{Var}(\hat{\mu}_x)$, $n \text{Var}(\hat{\sigma}_x)$, $n \text{Var}(\hat{\sigma}_m)$ and $n \text{Var}(\hat{\sigma}_p)$. Detailed derivation of maximum likelihood estimates and the Fisher information matrix.

REFERENCES

- CHAPMAN, D. G. (1956). Estimating the parameters of a truncated gamma distribution. *Ann. Math. Statist.* **27** 498–506. [MR0078622](#)
- DORFMAN, R. (1943). The detection of defective members of large populations. *Ann. Math. Statist.* **14** 436–440.
- FARAGGI, D., REISER, B. and SCHISTERMAN, E. F. (2003). ROC curve analysis for biomarkers based on pooled assessments. *Stat. Med.* **22** 2515–2527.
- GUPTA, A. K. (1952). Estimation of the mean and standard deviation of a normal population from a censored sample. *Biometrika* **39** 260–273. [MR0051483](#)
- LIU, A. and SCHISTERMAN, E. F. (2003). Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biom. J.* **45** 631–644. [MR1998141](#)
- LIU, A., SCHISTERMAN, E. F. and TEOH, E. (2004). Sample size and power calculation in comparing diagnostic accuracy of biomarkers with pooled assessments. *J. Appl. Stat.* **31** 49–59. [MR2041555](#)
- LOUIS, G., WEINER, J., WHITCOMB, B., SPERRAZZA, R., SCHISTERMAN, E., LOBDELL, D., CRICKARD, K., GREIZERSTEIN, H. and KOSTYNIK, P. (2005). Environmental PCB exposure and risk of endometriosis. *Human Reproduction* **20** 279–285.
- MUMFORD, S. L., SCHISTERMAN, E. F., VEXLER, A. and LIU, A. (2006). Pooling biospecimens and limits of detection: Effects on ROC curve analysis. *Biostatistics* **7** 585–598.
- RICHARDSON, D. B. and CIAMPI, A. (2003). Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *Am. J. Epidemiol.* **157** 355–363.
- SCHISTERMAN, E. and VEXLER, A. (2008). To pool or not to pool, from whether to when: Applications of pooling to biospecimens subject to a limit of detection. *Paediatric and Perinatal Epidemiology* **22** 486–496.
- SCHISTERMAN, E., FARAGGI, D., REISER, B. and TREVISAN, M. (2001). Statistical inference for the area under the receiver operating characteristic curve in the presence of random measurement error. *Am. J. Epidemiol.* **154** 174–179.
- SCHISTERMAN, E. F., PERKINS, N. J., LIU, A. and BONDELL, H. (2005). Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. *Epidemiology* **16** 73–81.
- SCHISTERMAN, E. F., VEXLER, A., WHITCOMB, B. W. and LIU, A. (2006). The limitations due to exposure detection limits for regression models. *Am. J. Epidemiol.* **163** 374–383.

- SCHISTERMAN, E. F., VEXLER, A., MUMFORD, S. L. and PERKINS, N. J. (2010). Hybrid pooled-unpooled design for cost-efficient measurement of biomarkers. *Stat. Med.* **29** 597–613. [MR2758456](#)
- SCHISTERMAN, E. F., VEXLER, A., YE, A. and PERKINS, N. J. (2011). Supplement to “A combined efficient design for biomarker data subject to a limit of detection due to measuring instrument sensitivity.” [DOI:10.1214/11-AOAS490SUPP](#).
- SHAM, P., BADER, J. S., CRAIG, I., O’DONOVAN, M. and OWEN, M. (2002). DNA pooling: A tool for large-scale association studies. *Nature Reviews Genetics* **3** 862–871.
- VEXLER, A., LIU, A. and SCHISTERMAN, E. F. (2006). Efficient design and analysis of biospecimens with measurements subject to detection limit. *Biom. J.* **48** 780–791. [MR2291289](#)
- VEXLER, A., SCHISTERMAN, E. F. and LIU, A. (2008). Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures. *Stat. Med.* **27** 280–296. [MR2412708](#)
- VEXLER, A., LIU, A., ELISEEVA, E. and SCHISTERMAN, E. F. (2008). Maximum likelihood ratio tests for comparing the discriminatory ability of biomarkers subject to limit of detection. *Biometrics* **64** 895–903. [MR2526641](#)
- WEINBERG, C. and UMBACH, D. (1999). Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics* **55** 718–726.
- YEE, T. W. (2010). VGAM: Vector generalized linear and additive models. R package version 0.8-1.
- ZHANG, S.-D. and GANT, T. W. (2005). Effect of pooling samples on the efficiency of comparative studies using microarrays. *Bioinformatics* **21** 4378–4383.

E. F. SCHISTERMAN

A. YE

N. J. PERKINS

EUNICE KENNEDY SHRIVER NATIONAL INSTITUTE
OF CHILD HEALTH AND HUMAN DEVELOPMENT

6100 EXECUTIVE BOULEVARD

ROCKVILLE, MARYLAND 20852

USA

E-MAIL: schistee@mail.nih.gov

yea2@mail.nih.gov

perkinsn@mail.nih.gov

A. VEXLER

UNIVERSITY OF NEW YORK AT BUFFALO

246 FARBER HALL

3435 MAIN STREET

BUFFALO, NEW YORK 14214

USA

E-MAIL: avexler@buffalo.edu